PNAS
www.pnas.org

Supplementary Information for

Benign Overfitting in Linear Regression

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, Alexander Tsigler

Peter Bartlett. Email: peter@berkeley.edu

This PDF file includes:

Supporting Information Text

A. Proof of Lemma 2. We first give the decomposition of the excess risk.

Lemma S.1. *The excess risk of the minimum norm estimator satisfies*

$$R(\hat{\theta}) = \mathbb{E}_x \left(x^\top \left(\theta^* - \hat{\theta} \right) \right)^2 \leq 2\theta^{*\top} B \theta^* + 2\boldsymbol{\varepsilon}^\top C \boldsymbol{\varepsilon},$$

and

$$\mathbb{E}_{x,\boldsymbol{\varepsilon}} R(\hat{\theta}) \geq \theta^{*\top} B \theta^* + \sigma^2 \operatorname{tr}(C),$$

where

$$\begin{aligned} B &= \left(I - X^\top \left(X X^\top \right)^{-1} X \right) \Sigma \left(I - X^\top \left(X X^\top \right)^{-1} X \right), \\ C &= \left(X X^\top \right)^{-1} X \Sigma X^\top \left(X X^\top \right)^{-1}. \end{aligned}$$

Proof. Since $\varepsilon = y - x^\top \theta^*$ has mean zero conditionally on x,

$$\begin{aligned} R(\hat{\theta}) &= \mathbb{E}_{x,y} \left(y - x^\top \hat{\theta} \right)^2 - \mathbb{E} \left(y - x^\top \theta^* \right)^2 \\ &= \mathbb{E}_{x,y} \left(y - x^\top \theta^* + x^\top \left(\theta^* - \hat{\theta} \right) \right)^2 - \mathbb{E} \left(y - x^\top \theta^* \right)^2 \\ &= \mathbb{E}_x \left(x^\top \left(\theta^* - \hat{\theta} \right) \right)^2. \end{aligned}$$

Using Eq. [1], the definition of Σ, and the fact that $\boldsymbol{y} = X\theta^* + \boldsymbol{\varepsilon}$,

$$\begin{aligned} R(\hat{\theta}) &= \mathbb{E}_x \left(x^\top \left(I - X^\top \left(X X^\top \right)^{-1} X \right) \theta^* - x^\top X^\top \left(X X^\top \right)^{-1} \boldsymbol{\varepsilon} \right)^2 \\ &\leq 2\mathbb{E}_x \left(x^\top \left(I - X^\top \left(X X^\top \right)^{-1} X \right) \theta^* \right)^2 + 2\mathbb{E}_x \left(x^\top X^\top \left(X X^\top \right)^{-1} \boldsymbol{\varepsilon} \right)^2 \\ &= 2\theta^{*\top} \left(I - X^\top \left(X X^\top \right)^{-1} X \right) \Sigma \left(I - X^\top \left(X X^\top \right)^{-1} X \right) \theta^* + 2\boldsymbol{\varepsilon}^\top \left(X X^\top \right)^{-1} X \Sigma X^\top \left(X X^\top \right)^{-1} \boldsymbol{\varepsilon} \\ &= 2\theta^{*\top} B \theta^* + 2\boldsymbol{\varepsilon}^\top C \boldsymbol{\varepsilon}. \end{aligned}$$

Also, since $\boldsymbol{\varepsilon}$ has zero mean conditionally on X, and is independent of x, we have

$$\begin{aligned} \mathbb{E}_{x,\boldsymbol{\varepsilon}} R(\hat{\theta}) &= \mathbb{E}_{x,\boldsymbol{\varepsilon}} \left[\left(x^\top \left(I - X^\top \left(X X^\top \right)^{-1} X \right) \theta^* \right)^2 + \left(x^\top X^\top \left(X X^\top \right)^{-1} \boldsymbol{\varepsilon} \right)^2 \right] \\ &= \theta^{*\top} \left(I - X^\top \left(X X^\top \right)^{-1} X \right) \Sigma \left(I - X^\top \left(X X^\top \right)^{-1} X \right) \theta^* + \operatorname{tr} \left(\left(X X^\top \right)^{-1} X \Sigma X^\top \left(X X^\top \right)^{-1} \mathbb{E} \left[\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top \middle| X \right] \right) \\ &\geq \theta^{*\top} B \theta^* + \sigma^2 \operatorname{tr}(C). \end{aligned}$$

□

The following lemma shows that we can obtain a high-probability upper bound on the term $\boldsymbol{\varepsilon}^\top C \boldsymbol{\varepsilon}$ in terms of the trace of C. It is Lemma 36 in (1).

Lemma S.2. *Consider random variables $\varepsilon_1, \ldots, \varepsilon_n$, conditionally independent given X and conditionally σ^2 sub-Gaussian, that is, for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[\exp(\lambda \varepsilon_i) | X] \leq \exp(\sigma^2 \lambda^2 / 2).$$

Suppose that, given X, $M \in \mathbb{R}^{n \times n}$ is a.s. positive semidefinite. Then a.s. on X, with conditional probability at least $1 - e^{-t}$,

$$\boldsymbol{\varepsilon}^\top M \boldsymbol{\varepsilon} \leq \sigma^2 \operatorname{tr}(M) + 2\sigma^2 \|M\| t + 2\sigma^2 \sqrt{\|M\|^2 t^2 + \operatorname{tr}\left(M^2\right) t}.$$

Since $\|C\| \leq \operatorname{tr}(C)$ and $\operatorname{tr}(C^2) \leq \operatorname{tr}(C)^2$, with probability at least $1 - e^{-t}$,

$$\boldsymbol{\varepsilon}^\top C \boldsymbol{\varepsilon} \leq \sigma^2 \operatorname{tr}(C)(2t+1) + 2\sigma^2 \sqrt{\operatorname{tr}(C)^2 (t^2 + t)} \leq (4t+2)\sigma^2 \operatorname{tr}(C).$$

Combining this with Lemma S.1 implies Lemma 2.

B. An Algebraic Property. .

Lemma S.3. *Suppose $k < n$, $A \in \mathbb{R}^{n\times n}$ is an invertible matrix, and $Z \in \mathbb{R}^{n\times k}$ is such that $ZZ^\top + A$ is invertible. Then*

$$Z^\top (ZZ^\top + A)^{-2} Z = (I + Z^\top A^{-1} Z)^{-1} Z^\top A^{-2} Z (I + Z^\top A^{-1} Z)^{-1}.$$

Proof. We use the Sherman–Morrison–Woodbury formula to write

$$(ZZ^\top + A)^{-1} = A^{-1} - A^{-1} Z (I + Z^\top A^{-1} Z)^{-1} Z^\top A^{-1}. \tag{S1}$$

Denote $M_1 := Z^\top A^{-1} Z$ and $M_2 := Z^\top A^{-2} Z$. Applying Eq. [S1], we get

$$\begin{aligned}
Z^\top (ZZ^\top + A)^{-2} Z &= Z^\top \left(A^{-1} - A^{-1} Z (I + Z^\top A^{-1} Z)^{-1} Z^\top A^{-1}\right)^2 Z \\
&= Z^\top \left(A^{-1} - A^{-1} Z (I + M_1)^{-1} Z^\top A^{-1}\right)^2 Z \\
&= Z^\top \Big(A^{-2} - A^{-2} Z (I + M_1)^{-1} Z^\top A^{-1} - A^{-1} Z (I + M_1)^{-1} Z^\top A^{-2} \\
&\qquad + A^{-1} Z (I + M_1)^{-1} Z^\top A^{-2} Z (I + M_1)^{-1} Z^\top A^{-1}\Big) Z \\
&= M_2 - M_2 (I + M_1)^{-1} M_1 - M_1 (I + M_1)^{-1} M_2 + M_1 (I + M_1)^{-1} M_2 (I + M_1)^{-1} M_1 \\
&= M_2 - M_2 (I + M_1)^{-1} M_1 - M_1 (I + M_1)^{-1} M_2 (I - (I + M_1)^{-1} M_1) \\
&= M_2 (I + M_1)^{-1} - M_1 (I + M_1)^{-1} M_2 (I + M_1)^{-1} \\
&= (I + M_1)^{-1} M_2 (I + M_1)^{-1},
\end{aligned}$$

where we used the identity $I - (I+M_1)^{-1} M_1 = (I+M_1)^{-1}$ twice in the second last equality and the identity $I - M_1 (I+M_1)^{-1} = (I + M_1)^{-1}$ in the last equality. □

C. Proof of concentration inequalities.

We use some standard results about sub-Gaussian and subexponential random variables. First of all, we need the following direct consequence of Propositions 2.5.2 and 2.7.1 and Lemma 2.7.6 from (2):

Lemma S.4. *There is a universal constant c such that for any random variable ξ that is centered, σ^2 sub-Gaussian, and unit variance, $\xi^2 - 1$ is a centered $c\sigma^2$-subexponential random variable, that is,*

$$\mathbb{E} \exp(\lambda(\xi^2 - 1)) \le \exp(c^2 \sigma^4 \lambda^2) \text{ for all such } \lambda \text{ that } |\lambda| \le \frac{1}{c\sigma^2}.$$

Second, we are going to use the following form of Bernstein's inequality, which is Theorem 2.8.2 in (2):

Lemma S.5. *There is a universal constant c such that, for any independent, mean zero, σ-subexponential random variables $\xi_1, \dots, \xi_N$, any $a = (a_1, \dots, a_N) \in \mathbb{R}^n$, and any $t \ge 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^N a_i \xi_i\right| > t\right) \le 2 \exp\left[-c \min\left(\frac{t^2}{\sigma^2 \sum_{i=1}^N a_i^2}, \frac{t}{\sigma \max_{1\le i\le n} a_i}\right)\right].$$

Corollary S.6. *There is a universal constant c such that for any non-increasing sequence $\{\lambda_i\}_{i=1}^\infty$ of non-negative numbers such that $\sum_{i=1}^\infty \lambda_i < \infty$, and any independent, centered, σ-subexponential random variables $\{\xi_i\}_{i=1}^\infty$, and any $x > 0$, with probability at least $1 - 2e^{-x}$*

$$\left|\sum_i \lambda_i \xi_i\right| \le c\sigma \max\left(x\lambda_1, \sqrt{x \sum_i \lambda_i^2}\right).$$

Corollary S.7. *There is a universal constant c such that for any centered random vector $z \in \mathbb{R}^n$ with independent σ^2 sub-Gaussian coordinates with unit variances, any random subspace $\mathscr{L}$ of $\mathbb{R}^n$ of codimension k that is independent of z, and any $t > 0$, with probability at least $1 - 3e^{-t}$,*

$$\begin{aligned}
\|z\|^2 &\le n + c\sigma^2 (t + \sqrt{nt}), \\
\|\Pi_{\mathscr{L}} z\|^2 &\ge n - c\sigma^2 (k + t + \sqrt{nt}),
\end{aligned}$$

where $\Pi_{\mathscr{L}}$ is the orthogonal projection on $\mathscr{L}$.

Proof. First of all, since $\|z\|^2 = \sum_{i=1}^n z_i^2$ — a sum of n σ^2-subexponential random variables, by Corollary S.6, for some absolute constant c and for any $t > 0$, with probability at least $1 - 2e^{-t}$,

$$\left|\|z\|^2 - n\right| \le c\sigma^2 \max(t, \sqrt{nt}).$$

Second, we can write

$$\|\Pi_{\mathscr{L}} z\|^2 = \|z\|^2 - \|\Pi_{\mathscr{L}^\perp} z\|^2.$$

Denote $M = \Pi_{\mathscr{L}^\perp}^\top \Pi_{\mathscr{L}^\perp}$. Since $\|M\| = 1$ and $\operatorname{tr}(M) = \operatorname{tr}(M^2) = k$, by Lemma S.2, with probability at least $1 - e^{-t}$,

$$\begin{aligned}\|\Pi_{\mathscr{L}^\perp} z\|^2 =& z^\top M z \\ \le& \sigma^2 k + 2\sigma^2 t + 2\sigma^2\sqrt{t^2 + kt} \\ \le& \sigma^2(2k + 4t).\end{aligned}$$

Thus, with probability at least $1 - 3e^{-t}$

$$\begin{aligned}\|z\|^2 \le& n + c\sigma^2 \max(t, \sqrt{nt}), \\ \|\Pi_{\mathscr{L}} z\|^2 \ge& \|z\| - \sigma^2(2k + 4t) \\ \ge& n - \sigma^2(2k + 4t + c\max(t, \sqrt{nt})).\end{aligned}$$

□

Lemma S.8 (ϵ-net argument)**.** *Suppose $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix, and $\mathcal{N}_\epsilon$ is an ϵ-net on the unit sphere $\mathcal{S}^{n-1}$ in the Euclidean norm, where $\epsilon < \frac{1}{2}$. Then*

$$\|A\| \le (1 - \epsilon)^{-2} \max_{x \in \mathcal{N}_\epsilon} |x^\top A x|.$$

Proof. Denote the eigenvalues of A as $\lambda_1, \ldots, \lambda_n$ and assume $|\lambda_1| \ge |\lambda_2| \ge \cdots \ge |\lambda_n|$. Denote the first eigenvector of A as $v \in \mathcal{S}^{n-1}$, and take $\Delta v \in \mathbb{R}^n$ such that $v + \Delta v \in \mathcal{N}_\epsilon$ and $\|\Delta v\| \le \epsilon$. Denote the coordinates of Δv in the eigenbasis of A as $\Delta v_1, \ldots, \Delta v_n$. Now we can write

$$\begin{aligned}\left|(v + \Delta v)^\top A (v + \Delta v)\right| &= \left|\lambda_1 + 2\lambda_1 \Delta v_1 + \sum_{i=1}^n \lambda_i \Delta v_i^2\right| \\ &= |\lambda_1| \cdot \left|1 + 2\Delta v_1 + \Delta v_1^2 + \sum_{i=2}^n \frac{\lambda_i}{\lambda_1} \Delta v_i^2\right| \\ &\ge |\lambda_1| \cdot \left|1 + 2\Delta v_1 + \Delta v_1^2 - \sum_{i=2}^n \Delta v_i^2\right| \\ &= |\lambda_1| \cdot \left|1 + 2\Delta v_1 + \Delta v_1^2 - \|\Delta v\|^2 + \Delta v_1^2\right| \\ &= |\lambda_1| \cdot \left|1 + 2\left(\Delta v_1 + \Delta v_1^2\right) - \|\Delta v\|^2\right| \\ &\ge |\lambda_1| \cdot \left|1 + 2\left(-\|\Delta v\| + (-\|\Delta v\|)^2\right) - \|\Delta v\|^2\right| \\ &= |\lambda_1| \cdot \left|1 - 2\|\Delta v\| + \|\Delta v\|^2\right| \\ &\ge |\lambda_1| \cdot |1 - 2\epsilon + \epsilon^2| \\ &= \|A\|(1 - \epsilon)^2,\end{aligned}$$

where the first inequality holds because the λ_is are decreasing in magnitude, and the last two inequalities hold since the functions $x + x^2$ and $2x + x^2$ are both increasing on $(-\frac{1}{2}, \infty)$ and $\Delta v_1 \ge -\|\Delta v\| \ge -\epsilon \ge -\frac{1}{2}$. □

We restate Lemma 4.

Lemma S.9. *There is a universal constant c such that with probability at least $1 - 2e^{-n/c}$,*

$$\frac{1}{c}\sum_i \lambda_i - c\lambda_1 n \le \mu_n(A) \le \mu_1(A) \le c\left(\sum_i \lambda_i + \lambda_1 n\right).$$

Proof. For a fixed vector $v \in \mathbb{R}^n$, Proposition 2.6.1 from (2) implies that for some constant c_1 and any i the random variable $v^\top z_i$ is $c_1\|v\|^2\sigma_x^2$ sub-Gaussian. Thus, for any fixed unit vector v, as $v^\top A v = \sum_i \lambda_i (v^\top z_i)^2$, Lemma S.4 and Corollary S.6 imply that for some constant c_2 with probability at least $1 - 2e^{-t}$,

$$\left| v^\top A v - \sum \lambda_i \right| \le c_2 \sigma_x^2 \max\left(\lambda_1 t, \sqrt{t \sum \lambda_i^2} \right).$$

Let $\mathcal{N}$ be a $\frac{1}{4}$-net on the sphere $\mathcal{S}^{n-1}$ with respect to the Euclidean distance such that $|\mathcal{N}| \le 9^n$. Applying the union bound over the elements of $\mathcal{N}$, we see that with probability $1 - 2e^{-t}$, every $v \in \mathcal{N}$ satisfies

$$\left| v^\top A v - \sum \lambda_i \right| \le c_2 \sigma_x^2 \max\left(\lambda_1 (t + n \ln 9), \sqrt{(t + n \ln 9) \sum_i \lambda_i^2} \right).$$

Since $\mathcal{N}$ is a $\frac{1}{4}$-net, by Lemma S.8, we need to multiply the quantity above by $(1 - 1/4)^{-2}$ to get the bound on the norm of the $A - I_n \sum_i \lambda_i$. Denote

$$\Diamond = \left(\lambda_1 (t + n \ln 9) + \sqrt{(t + n \ln 9) \sum_i \lambda_i^2} \right).$$

Thus, with probability at least $1 - 2e^{-t}$,

$$\left\| A - I_n \sum_i \lambda_i \right\| \le c_3 \sigma_x^2 \Diamond.$$

When $t < n/c_4$ we can write $t + n \ln 9 \le c_5 n$, and we have

$$\begin{aligned}
\Diamond &\le c_5 \left(\lambda_1 n + \sqrt{n \sum_i \lambda_i^2} \right) \\
&\le c_5 \lambda_1 n + \sqrt{(c_5^2 \lambda_1 n) \sum_i \lambda_i} \\
&\le c_6 \sigma_x^2 \lambda_1 n + \frac{1}{2 c_3 \sigma_x^2} \sum_i \lambda_i,
\end{aligned}$$

by the AMGM inequality. (Recall that $c_1, c_2, \ldots$ denote universal constants with value at least 1, and $\sigma_x \ge 1/c_7$ is the sub-Gaussian constant of a random variable with unit variance.) □

D. Proof of Lemma 8. Fix $i \ge 1$ with $\lambda_i > 0$ and $0 \le k \le n/c$. By Lemma 5, with probability at least $1 - 2e^{-n/c_1}$,

$$\mu_{k+1}(A_{-i}) \le c_1 \left(\sum_{j>k} \lambda_j + \lambda_{k+1} n \right),$$

and hence

$$z_i^\top A_{-i}^{-1} z_i \ge \frac{\|\Pi_{\mathscr{L}_i} z_i\|^2}{c_1 \left(\sum_{j>k} \lambda_j + \lambda_{k+1} n \right)}.$$

By Corollary 1, with probability at least $1 - 3e^{-t}$,

$$\|\Pi_{\mathscr{L}_i} z_i\|^2 \ge n - a\sigma_x^2 (k + t + \sqrt{tn}) \ge n/c_2,$$

provided that $t < n/c_0$ and $c > c_0$ for some sufficiently large c_0. Thus, with probability at least $1 - 5e^{-n/c_3}$,

$$z_i^\top A_{-i}^{-1} z_i \ge \frac{n}{c_3 \left(\sum_{j>k} \lambda_j + \lambda_{k+1} n \right)},$$

hence

$$1 + \lambda_i z_i^\top A_{-i}^{-1} z_i \le \left(\frac{c_3 \left(\sum_{j>k} \lambda_j + \lambda_{k+1} n \right)}{\lambda_i n} + 1 \right) \lambda_i z_i^\top A_{-i}^{-1} z_i.$$

Dividing $\lambda_i^2 z_i^\top A_{-i}^{-2} z_i$ by the square of both sides, we have

$$\frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1+\lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \geq \left(\frac{c_3 \left(\sum_{j>k} \lambda_j + \lambda_{k+1} n \right)}{\lambda_i n} + 1 \right)^{-2} \frac{z_i^\top A_{-i}^{-2} z_i}{(z_i^\top A_{-i}^{-1} z_i)^2}.$$

Also, from the Cauchy-Schwarz inequality and Corollary 1 again, we have that on the same event,

$$\begin{aligned} \frac{z_i^\top A_{-i}^{-2} z_i}{(z_i^\top A_{-i}^{-1} z_i)^2} &\geq \frac{z_i^\top A_{-i}^{-2} z_i}{\left\| A_{-i}^{-1} z_i \right\|^2 \|z_i\|^2} \\ &= \frac{1}{\|z_i\|^2} \geq \frac{1}{n + a\sigma_x^2 (t + \sqrt{nt})} \geq \frac{1}{c_4 n}. \end{aligned}$$

Choosing c suitably large gives the lemma.

E. Proof of Lemma 9. We know that, for all $i \leq n$, $\mathbb{P}(\eta_i > t_i) \geq 1 - \delta$. Consider the following event:

$$E = \left\{ \sum_{i=1}^n \eta_i < \frac{1}{2} \sum_{i=1}^n t_i \right\},$$

and denote its probability as $c\delta$ for some $c \in (0, \delta^{-1})$. On the one hand, by the definition of the event, we have

$$\frac{1}{\mathbb{P}(E)} \mathbb{E} \left[\mathbb{1}_E \sum_{i=1}^n \eta_i \right] \leq \frac{1}{2} \sum_{i=1}^n t_i.$$

On the other hand, note that for any i,

$$\begin{aligned} \mathbb{E}[\eta_i \mathbb{1}_E] \geq & \mathbb{E}[t_i \mathbb{1}_{\{\eta_i \geq t_i\} \cap E}] \\ &= t_i \mathbb{P}(\{\eta_i \geq t_i\} \cap E) \\ &\geq t_i (\mathbb{P}\{\eta_i \geq t_i\} + \mathbb{P}(E) - 1) \\ &\geq t_i (c-1)\delta. \end{aligned}$$

So

$$\begin{aligned} \mathbb{E} \left[\mathbb{1}_E \sum_{i=1}^n \eta_i \right] &\geq (c-1)\delta \sum_{i=1}^n t_i, \\ \frac{1}{\mathbb{P}(E)} \mathbb{E} \left[\mathbb{1}_E \sum_{i=1}^n \eta_i \right] &\geq (1 - c^{-1}) \sum_{i=1}^n t_i. \end{aligned}$$

Thus, we obtain

$$\begin{aligned} \frac{1}{2} \sum_{i=1}^n t_i &\geq (1 - c^{-1}) \sum_{i=1}^n t_i, \\ c &\leq 2, \\ \mathbb{P} \left(\sum_{i=1}^n \eta_i < \frac{1}{2} \sum_{i=1}^n t_i \right) &= c\delta \leq 2\delta. \end{aligned}$$

F. Proof of Lemma 11. We can write the function of l being minimized as

$$\begin{aligned} \frac{l}{bn} + \frac{bn \sum_{i>l} \lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} &= \sum_{i=1}^l \frac{1}{bn} + \sum_{i>l} \frac{bn\lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} \\ &\geq \sum_{i=1}^{k^*} \min \left\{ \frac{1}{bn}, \frac{bn\lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} \right\} \\ &\quad + \sum_{i>k^*} \frac{bn\lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} \\ &= \sum_{i=1}^{l^*} \frac{1}{bn} + \sum_{i>l^*} \frac{bn\lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2}, \end{aligned}$$

where l^* is the largest value of $i \leq k^*$ for which

$$\frac{1}{bn} \leq \frac{bn\lambda_i^2}{\left(\lambda_{k^*+1} r_{k^*}(\Sigma)\right)^2},$$

since the λ_i^2 are non-increasing. This condition holds iff

$$\lambda_i \geq \frac{\lambda_{k^*+1} r_{k^*}(\Sigma)}{bn}.$$

The definition of k^* implies $r_{k^*-1}(\Sigma) < bn$. So we can write

$$\begin{aligned} r_{k^*}(\Sigma) &= \frac{\sum_{i>k^*} \lambda_i}{\lambda_{k^*+1}} \\ &= \frac{\sum_{i>k^*-1} \lambda_i - \lambda_{k^*}}{\lambda_{k^*+1}} \\ &= \frac{\lambda_{k^*}}{\lambda_{k^*+1}} (r_{k^*-1}(\Sigma) - 1) \\ &< \frac{\lambda_{k^*}}{\lambda_{k^*+1}} (bn - 1), \end{aligned}$$

and so the minimizing l is k^*. Also,

$$\frac{\sum_{i>k^*} \lambda_i^2}{\left(\lambda_{k^*+1} r_{k^*}(\Sigma)\right)^2} = \frac{\sum_{i>k^*} \lambda_i^2}{\left(\sum_{i>k^*} \lambda_i\right)^2} = \frac{1}{R_{k^*}(\Sigma)}.$$

G. Eigenvalue monotonicity. Recall (half of) the Courant-Fischer-Weyl theorem.

Lemma S.10. *For any symmetric $n \times n$ matrix A, and any $i \in [n]$, $\mu_i(A)$ is the minimum, over all subspaces U of $\mathbb{R}^n$ of dimension $n - i$, of the maximum, over all unit-length $u \in U$, of $u^\top A u$.*

Lemma S.11 (Monotonicity of eigenvalues)**.** *If symmetric matrices A and B satisfy $A \preceq B$, then, for any $i \in [n]$, we have $\mu_i(A) \leq \mu_i(B)$.*

Proof. Let U be the subspace of $\mathbb{R}^n$ of dimension $n - i$ that minimizes the maximum over all unit-length $u \in U$, of $u^\top A u$, and let V be the analogous subspace for B. We have

$$\begin{aligned} \mu_i(A) &= \max_{u \in U : ||u||=1} u^\top A u \quad \text{(by Lemma S.10)} \\ &\leq \max_{v \in V : ||v||=1} v^\top A v \quad \text{(since } U \text{ is the minimizer)} \\ &\leq \max_{v \in V : ||v||=1} v^\top B v \quad \text{(since } A \preceq B) \\ &= \mu_i(B), \end{aligned}$$

by Lemma S.10, completing the proof. □

H. Rank facts. The quantity $r_0(\Sigma)$ is an important complexity parameter for covariance estimation problems, where it has been called the 'effective rank' (3, 4). Earlier, $r_0(\Sigma^2)$ was called the 'stable rank' (5) and the 'numerical rank' (6), although that term has a different meaning in computational linear algebra (7, p261).

We restate Lemma 1.

Lemma S.12. *$r_k(\Sigma) \geq 1$, $r_k^2(\Sigma) = r_k(\Sigma^2) R_k(\Sigma)$, and $r_k(\Sigma^2) \leq r_k(\Sigma) \leq R_k(\Sigma) \leq r_k^2(\Sigma)$.*

Proof. The first inequality and the equality are immediate from the definitions. Together they imply $R_k(\Sigma) \leq r_k^2(\Sigma)$. For the second inequality,

$$r_k(\Sigma^2) = \frac{\sum_{i>k} \lambda_i^2}{\lambda_{k+1}^2} \leq \frac{\lambda_{k+1} \sum_{i>k} \lambda_i}{\lambda_{k+1}^2} = r_k(\Sigma).$$

Substituting this in the equality implies $r_k(\Sigma) \leq R_k(\Sigma)$. □

Lemma S.13. *Writing r_k and R_k for $r_k(\Sigma)$ and $R_k(\Sigma)$,*

$$\frac{1}{R_{k+1}} = \frac{\frac{1}{R_k} - \frac{1}{r_k^2}}{1 - \left(2 - \frac{1}{r_k}\right) \frac{1}{r_k}}.$$

Thus, the function $\phi(k) = k/(b^2 n) + n/R_k$ satisfies the monotonicity property $\phi(k+1) > \phi(k)$ whenever $r_k > bn \geq 1$.

Proof. Writing

$$q = \sum_{i>k+1} \lambda_i^2, \quad s = \sum_{i>k+1} \lambda_i,$$

so that $R_{k+1} = s^2/q$, we have

$$\begin{aligned}
\frac{1}{R_k} - \frac{1}{R_{k+1}} &= \frac{\lambda_{k+1}^2 + q}{(\lambda_{k+1}+s)^2} - \frac{q}{s^2} \\
&= \frac{\left(\lambda_{k+1}^2 + q\right) s^2 - q\left(\lambda_{k+1}+s\right)^2}{s^2\left(\lambda_{k+1}+s\right)^2} \\
&= \frac{1}{r_k^2} - \frac{q\lambda_{k+1}\left(\lambda_{k+1}+2s\right)}{s^2\left(\lambda_{k+1}+s\right)^2} \\
&= \frac{1}{r_k^2} - \frac{2\left(\lambda_{k+1}+s\right) - \lambda_{k+1}}{R_{k+1} r_k \left(\lambda_{k+1}+s\right)} \\
&= \frac{1}{r_k^2} - \frac{2 - 1/r_k}{R_{k+1} r_k}.
\end{aligned}$$

Hence

$$\frac{1}{R_{k+1}} = \frac{1/R_k - 1/r_k^2}{1 - \left(2 - \frac{1}{r_k}\right)\frac{1}{r_k}}.$$

Since $r_k > 1$, $0 < 1 - (2 - 1/r_k)/r_k < 1$, so

$$\frac{n}{R_{k+1}} > \frac{n}{R_k} - \frac{n}{r_k^2},$$

and if $r_k > bn$,

$$\begin{aligned}
\phi(k+1) - \phi(k) &= \frac{k+1}{b^2 n} + \frac{n}{R_{k+1}} - \left(\frac{k}{b^2 n} + \frac{n}{R_k}\right) \\
&> \frac{1}{b^2 n} - \frac{n}{r_k^2} \\
&> 0.
\end{aligned}$$

□

I. Conditions on eigenvalues. In this section, we prove the following expanded version of Theorem 2.

Theorem S.14. *Define $\lambda_{k,n} := \mu_k(\Sigma_n)$ for all k, n.*

1. *If $\lambda_{k,n} = k^{-\alpha} \ln^{-\beta}((k+1)e/2)$, then Σ_n is benign iff $\alpha = 1$ and $\beta > 1$.*
2. *If $\lambda_{k,n} = k^{-(1+\alpha_n)}$, then Σ_n is benign iff $\omega(1/n) = \alpha_n = o(1)$. Furthermore,*

$$R(\hat{\theta}) = \Theta\left(\min\left\{\frac{1}{\alpha_n n} + \alpha_n, 1\right\}\right).$$

3. *If*

$$\lambda_{k,n} = \begin{cases} k^{-\alpha} & \text{if } k \le p_n, \\ 0 & \text{otherwise,} \end{cases}$$

then Σ_n is benign iff either $0 < \alpha < 1$, $p_n = \omega(n)$ and $p_n = o\left(n^{1/(1-\alpha)}\right)$ or $\alpha = 1$, $p_n = e^{\omega(\sqrt{n})}$ and $p_n = e^{o(n)}$.

4. *If*

$$\lambda_{k,n} = \begin{cases} \gamma_k + \epsilon_n & \text{if } k \le p_n, \\ 0 & \text{otherwise,} \end{cases}$$

and $\gamma_k = \Theta(\exp(-k/\tau))$, then Σ_n with $||\Sigma_n|| = 1$ is benign iff $p_n = \omega(n)$ and $ne^{-o(n)} = \epsilon_n p_n = o(n)$. Furthermore, for $p_n = \Omega(n)$ and $\epsilon_n p_n = ne^{-o(n)}$,

$$R(\hat{\theta}) = O\left(\frac{\epsilon_n p_n + 1}{n} + \frac{\ln(n/(\epsilon_n p_n))}{n} + \max\left\{\frac{1}{n}, \frac{n}{p_n}\right\}\right).$$

5. *If*

$$\lambda_{k,n} = \begin{cases} 1 & \text{if } k \le s_n, \\ \epsilon_n & \text{if } s_n < k \le p_n, \\ 0 & \text{otherwise,} \end{cases}$$

with $\epsilon_n > 0$, then Σ_n is benign iff $s_n = o(n)$, $p_n = \omega(n)$ and $\epsilon_n p_n = o(n)$.

6. *If*

$$\lambda_{k,n} = \begin{cases} 1 & \text{if } k = 1, \\ \epsilon_n \dfrac{1+\theta^2 - 2\theta\cos(k\pi/(p_n+1))}{1+\theta^2-2\theta\cos(\pi/(p_n+1))} & \text{if } 1 < k \le p_n, \\ 0 & \text{otherwise,} \end{cases}$$

with $\theta < 1$ and $\epsilon_n > 0$, then Σ_n is benign iff $p_n = \omega(n)$ and $\epsilon_n p_n = o(n)$.

Theorem S.14(5) shows that the isotropic case is not benign (with $s_n = p_n$) and that the equicorrelation case is benign (with $s_n = 1$), provided that the correlation coefficient is sufficiently small. Theorem S.14(6) shows that the covariance matrix for a length p_n sample path from an MA(1) time series model (see, for example, (8, 9)) is not benign (with $\epsilon_n = 1$), and that a spiked MA(1) model is benign, provided the MA(1) variance, ϵ_n, is sufficiently small.

We build up the proof in stages. First, we characterize those sequences of effective ranks that can arise.

Theorem S.15. *Consider some positive summable sequence $\{\lambda_i\}_{i=1}^{\infty}$, and for any non-negative integer i denote*

$$r_i := \lambda_{i+1}^{-1} \sum_{j>i} \lambda_j.$$

Then $r_i > 1$ and $\sum_i r_i^{-1} = \infty$. Moreover, for any positive sequence $\{u_i\}$ such that $\sum_{i=0}^{\infty} u_i^{-1} = \infty$ and for every i $u_i > 1$, there exists a positive sequence $\{\lambda_i\}$ (unique up to constant multiplier) such that $r_i \equiv u_i$. The sequence is (a constant rescaling of)

$$\lambda_k = u_{k-1}^{-1} \prod_{i=0}^{k-2} (1 - u_i^{-1}).$$

Proof.

$$\sum_{i \ge k+1} \lambda_i = \sum_{i \ge k} \lambda_i - \lambda_k = (1 - r_{k-1}^{-1}) \sum_{i \ge k} \lambda_i.$$

Thus,

$$\sum_{i \ge k+1} \lambda_i = \prod_{i=0}^{k-1} \left(1 - r_i^{-1}\right) \cdot \sum_i \lambda_i,$$

which goes to zero if and only if $\sum_i r_i^{-1} = \infty$. On the other hand, we may rewrite the first equality in the proof as

$$\lambda_{k+1} r_k = \lambda_k r_{k-1} (1 - r_{k-1}^{-1}),$$

and hence

$$\lambda_k r_{k-1} = \prod_{i=0}^{k-2} \left(1 - r_i^{-1}\right) \lambda_1 r_0.$$

So for any sequence $\{u_i\}$ we can uniquely (up to a constant multiplier) recover the sequence $\{\lambda_i\}$ such that $r_i = u_i$ — the only candidate is

$$\lambda_k = u_{k-1}^{-1} \prod_{i=0}^{k-2} (1 - u_i^{-1}).$$

However, for such $\{\lambda_i\}$ one can compute

$$\sum_{i=1}^{k} \lambda_i = 1 - \prod_{i=0}^{k-1} (1 - u_i^{-1}),$$

so the resulting sequence $\{\lambda_i\}$ sums to 1, and

$$r_k = \lambda_{k+1}^{-1} \sum_{i>k} \lambda_i = \lambda_{k+1}^{-1} \prod_{i=0}^{k-1} (1 - u_i^{-1}) = u_k.$$

□

Theorem S.16. *Suppose b is some constant, and $k^*(n) = \min\{k : r_k \geq bn\}$. Suppose also that the sequence $\{r_n\}$ is increasing. Then, as n goes to infinity, $k^*(n)/n$ goes to zero if and only if r_n/n goes to infinity.*

Proof. We prove the "if" part separately from the "only if" part.

1. **If $k^*(n)/n \to 0$ then $r_n/n \to \infty$.**

   Fix some $C > 1$. Since $k^*(n)/n \to 0$, there exists some N_C such that for any $n \geq N_C$, $k^*(n) < n/C$. Thus, for all $n > N_C$,

   $$k^*(\lfloor Cn \rfloor) \leq n,$$
   $$r_n \geq r_{k^*(\lfloor Cn \rfloor)} \geq b\lfloor Cn \rfloor.$$

   Since the constant C is arbitrary, r_n/n goes to infinity.

2. **If $r_n/n \to \infty$ then $k^*(n)/n \to 0$.**

   Fix some constant $C > 1$. Since $r_n/n \to \infty$ there exists some N_C such that for any $n \geq N_C$, $r_n > Cn$. Thus, for any $n > CN_C/b$

   $$r_{\lceil nb/C \rceil} \geq bn,$$
   $$k^*(n) \leq \lceil nb/C \rceil.$$

   Since the constant C is arbitrary, $k^*(n)/n$ goes to zero.

□

Theorem S.17. *Suppose the sequence $\{r_i\}$ is increasing and $r_n/n \to \infty$ as $n \to \infty$. Then a sufficient condition for $\frac{n}{R_{k^*(n)}} \to 0$ is*

$$r_k^{-2} = o(r_k^{-1} - r_{k+1}^{-1}) \text{ as } k \to \infty.$$

For example, this condition holds for $r_n = n \log n$.

Proof. We need to show that

$$\frac{n}{R_{k^*(n)}} = \frac{n \sum_{i>k^*(n)} \lambda_i^2}{\left(\sum_{i>k^*(n)} \lambda_i\right)^2} = \frac{n \sum_{i>k^*(n)} \lambda_i^2}{\lambda_{k^*(n)+1}^2 r_{k^*(n)}^2} \to 0.$$

Since $r_{k^*(n)} \geq bn$ and $\lim_{n\to\infty} k^*(n) = \infty$, it is enough to prove that $\frac{\sum_{i>k} \lambda_i^2}{\lambda_{k+1}^2 r_k} \to 0$ as k goes to infinity. Since

$$\lambda_{k+2} r_{k+1} = \lambda_{k+1} r_k (1 - r_k^{-1}),$$

we can write that

$$\begin{aligned}\lambda_{k+1+l} r_{k+l} &= \lambda_{k+1} r_k \prod_{i=k}^{k+l-1} (1 - r_i^{-1}) \\ &\leq \lambda_{k+1} r_k \exp\left(-\sum_{i=k}^{k+l-1} r_i^{-1}\right)\end{aligned}$$

which yields

$$\frac{\lambda_{k+1+l}}{\lambda_{k+1} r_k} \leq r_{k+l}^{-1} \exp\left(-\sum_{i=k}^{k+l-1} r_i^{-1}\right).$$

Thus, we obtain

$$\frac{\sum_{i>k} \lambda_i^2}{\lambda_{k+1}^2 r_k} \leq r_k \sum_{i\geq k} r_i^{-2} \exp\left(-2\sum_{j=k}^{i-1} r_j^{-1}\right),$$

and it is sufficient to prove that the latter quantity goes to zero. We write

$$\begin{aligned}r_k \sum_{i\geq k} r_i^{-2} \exp\left(-2\sum_{j=k}^{i-1} r_j^{-1}\right) &= \frac{\sum_{i\geq k} r_i^{-2} \exp\left(-2\sum_{j=k}^{i-1} r_j^{-1}\right)}{r_k^{-1}} \\ &= \frac{\sum_{i\geq k} r_i^{-2} \exp\left(-2\sum_{j=0}^{i-1} r_j^{-1}\right)}{r_k^{-1} \exp\left(-2\sum_{j=0}^{k-1} r_j^{-1}\right)}.\end{aligned}$$

Since both numerator and denominator are decreasing in k and go to zero as $k \to \infty$, we can apply the Stolz–Cesáro theorem (an analog of L'Hôpital's rule for discrete sequences):

$$\begin{aligned}
\lim_{k\to\infty} \frac{\sum_{i\geq k} r_i^{-2} \exp\left(-2\sum_{j=0}^{i-1} r_j^{-1}\right)}{r_k^{-1} \exp\left(-2\sum_{j=0}^{k-1} r_j^{-1}\right)} &= \lim_{k\to\infty} \frac{r_k^{-2} \exp\left(-2\sum_{j=0}^{k-1} r_j^{-1}\right)}{(r_k^{-1} - e^{-2r_k^{-1}} r_{k+1}^{-1}) \exp\left(-2\sum_{j=0}^{k-1} r_j^{-1}\right)} \\
&= \lim_{k\to\infty} \frac{r_k^{-2}}{(r_k^{-1} - e^{-2r_k^{-1}} r_{k+1}^{-1})} \qquad \text{(since, for large enough } k,\ e^{-2r_k^{-1}} \leq 1 - r_k^{-1}) \\
&\leq \lim_{k\to\infty} \frac{r_k^{-2}}{r_k^{-1} - r_{k+1}^{-1} + r_k^{-1} r_{k+1}^{-1}} \\
&= 0,
\end{aligned}$$

where the last line is due to our sufficient condition. □

Now we are ready to prove Theorem S.14.

Part 1, if direction, first term: We have

$$r_0(\Sigma_n) = \sum_{i=1}^{\infty} \lambda_i = O\left(\sum_{i=1}^{\infty} \frac{1}{i \log^{\beta}(1+i)}\right),$$

which is $O(1)$ for $\beta > 1$.

Part 1, if direction, second term: By Theorem S.16, it suffices to prove that $\lim_{n\to\infty} \frac{r_n}{n} = \infty$. This holds because

$$r_n = \frac{\sum_{i>n} \frac{1}{i \log^{\beta}(1+i)}}{\frac{1}{(n+1)\log^{\beta}(2+n)}} = \Theta(n \log n),$$

since $\beta > 1$.

Part 1, if direction, third term: By Theorem S.17, it suffices to prove that $r_k^{-2} = o(r_k^{-1} - r_{k+1}^{-1})$, that is

$$\lim_{k\to\infty} \frac{r_k^{-2}}{r_k^{-1} - r_{k+1}^{-1}} = 0$$

or, equivalently,

$$\lim_{k\to\infty} \frac{r_{k+1}}{r_k(r_{k+1} - r_k)} = 0.$$

As argued above, when $\alpha = 1$ and $\beta > 1$, $r_k = \Theta(k \log k)$, so it suffices to show that $\lim_{k\to\infty}(r_{k+1} - r_k) = \infty$. We have

$$\begin{aligned}
r_{k+1} - r_k &= \frac{\sum_{i>k+1} \lambda_i}{\lambda_{k+2}} - \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}} \\
&= \frac{\left((\lambda_{k+1} - \lambda_{k+2}) \sum_{i>k+1} \lambda_i\right) - \lambda_{k+1}\lambda_{k+2}}{\lambda_{k+1}\lambda_{k+2}} \\
&= \left(\left(\frac{1}{\lambda_{k+2}} - \frac{1}{\lambda_{k+1}}\right) \sum_{i>k+1} \lambda_i\right) - 1
\end{aligned}$$

so it suffices to show that

$$\lim_{k\to\infty} \left(\frac{1}{\lambda_{k+2}} - \frac{1}{\lambda_{k+1}}\right) \sum_{i>k+1} \lambda_i = \infty.$$

Since λ_i is non-increasing, we have

$$\begin{aligned}
\left(\frac{1}{\lambda_{k+2}} - \frac{1}{\lambda_{k+1}}\right) \sum_{i>k+1} \lambda_i &\geq \left(\frac{1}{\lambda_{k+2}} - \frac{1}{\lambda_{k+1}}\right) \int_{k+1}^{\infty} \frac{1}{x \log^{\beta} x}\, dx \\
&= \left(\frac{1}{\lambda_{k+2}} - \frac{1}{\lambda_{k+1}}\right) \frac{1}{(\beta-1)\log^{\beta-1}(k+1)} \\
&= \frac{(k+2)\log^{\beta}(k+3) - (k+1)\log^{\beta}(k+2)}{(\beta-1)\log^{\beta-1}(k+1)}.
\end{aligned}$$

If we define f on the positive reals by $f(x) = x\log^\beta(x+1)$, then f is convex, and, since $f'(x) = \frac{\beta x \log^{\beta-1}(x+1)}{x+1} + \log^\beta(x+1)$, we have

$$\frac{(k+2)\log^\beta(k+3) - (k+1)\log^\beta(k+2)}{(\beta-1)\log^{\beta-1}(k+1)} \geq \frac{\frac{\beta(k+1)\log^{\beta-1}(k+2)}{k+2} + \log^\beta(k+2)}{(\beta-1)\log^{\beta-1}(k+1)},$$

which goes to infinity for large k, completing the proof of the "if" direction of the third term of Part 1.

Part 1, only if direction, $\alpha > 1$: If $\alpha > 1$, then

$$\begin{aligned} r_n &= \frac{\sum_{i>n} \frac{1}{i^a \log^\beta(1+i)}}{\frac{1}{n^a \log^\beta(1+n)}} \\ &\leq n^\alpha \sum_{i>n} \frac{\log^\beta(1+n)}{i^a \log^\beta(1+i)} \\ &\leq n^\alpha \sum_{i>n} \frac{1}{i^a} \\ &= n^\alpha O(n^{1-\alpha}), \end{aligned}$$

which does not grow faster than n. Thus, by Theorem S.16, $k^*(n)/n$ does not go to zero.

Part 1, only if direction, $\alpha < 1$, or $\alpha = 1$ and $\beta \leq 1$: In this case, since, as above

$$r_0(\Sigma_n) = \sum_{i=1}^\infty \lambda_i,$$

and $\sum_{i=1}^\infty \frac{1}{i^\alpha \log^\beta(1+i)}$ diverges in this case, $\frac{r_0(\Sigma_n)}{n}$ does not go to zero.

Before starting on Part 2, let us define $r_{k,n} = r_k(\Sigma_n)$ and $R_{k,n} = R_k(\Sigma_n)$.

Part 2, if direction, first term: We have

$$r_{0,n} = \sum_{i=1}^\infty \lambda_{i.n} = \sum_{i=1}^\infty \frac{1}{i^{1+\alpha_n}} \leq 1 + \frac{1}{\alpha_n},$$

so $\frac{r_{0,n}}{n} \leq \frac{1+\frac{1}{\alpha_n}}{n}$ which goes to zero with n if $\alpha_n = \omega(1/n)$.

Part 2, if direction, second term: First,

$$\begin{aligned} r_{k,n} &= (k+1)^{1+\alpha_n} \sum_{i>k} i^{-(1+\alpha_n)} \\ &\geq (k+1)^{1+\alpha_n} \int_{k+1}^\infty x^{-(1+\alpha_n)} dx \\ &= \frac{k+1}{\alpha_n}. \end{aligned}$$

Thus, $k^*(n) = O(\alpha_n n)$, so that $\frac{k^*(n)}{n} = O(\alpha_n) = o(1)$.

Part 2, if direction, third term: We bound $R_{k,n}$ from below by separately bounding its numerator and denominator:

$$\begin{aligned} \sum_{i>k} i^{-(1+\alpha_n)} &\geq \int_{k+1}^\infty x^{-(1+\alpha_n)}\, dx \\ &= \frac{1}{\alpha_n (k+1)^{\alpha_n}}, \end{aligned}$$

and

$$\begin{aligned} \sum_{i>k} i^{-2(1+\alpha_n)} &\leq \int_{k}^\infty x^{-2(1+\alpha_n)}\, dx \\ &= \frac{1}{k^{1+2\alpha_n}(2\alpha_n+1)}, \end{aligned}$$

so that

$$R_{k,n} \geq \frac{k^{1+2\alpha_n}(2\alpha_n+1)}{\alpha_n^2 (k+1)^{2\alpha_n}} \geq \frac{k}{\alpha_n^2} \times \left(1 - \frac{1}{k+1}\right)^{2\alpha_n}. \qquad [S2]$$

So now we want a lower bound on $k^*(n)$. For that, we need an upper bound on $r_{k,n}$, and

$$\begin{aligned} r_{k,n} &\leq (k+1)^{1+\alpha_n} \int_k^\infty x^{-(1+\alpha_n)} dx \\ &= \frac{(k+1)}{\alpha_n} \times \left(1+\frac{1}{k}\right)^{\alpha_n} \\ &\leq \frac{2k}{\alpha_n} e^{\alpha_n/k}. \end{aligned}$$

This implies $\frac{2k^*(n)}{\alpha_n} e^{\alpha_n/k^*(n)} \geq bn$. This, together with the fact that, for $u > 1$, $ue^{1/u}$ is an increasing function of u, implies that, for large enough n, $k^*(n) \geq \alpha_n bn/3$. Since $\alpha_n = \omega(1/n)$, this implies that $k^*(n) = \omega(1)$. Combining this with Eq. [S2], for large enough n

$$R_{k^*(n),n} \geq \frac{k^*(n)}{\alpha_n^2} e^{-\alpha_n/k^*(n)} \geq \frac{k^*(n)}{2\alpha_n^2} \geq \frac{bn}{6\alpha_n}.$$

Thus $n/R_{k^*(n),n} = O(\alpha_n) = o(1)$.

Part 2, only if direction, $\alpha_n = O(1/n)$**:** We have

$$r_{0,n} = \sum_{i=1}^\infty \frac{1}{i^{1+\alpha_n}} \geq \frac{1}{\alpha_n},$$

so $\frac{r_{0,n}}{n} \geq \frac{1}{\alpha_n n}$, which is bounded below by a constant for large n if $\alpha_n = O(1/n)$.

Part 2, only if direction, $\alpha_n = \Omega(1)$**:** Recall that, in the proof of the "if" direction of the third term, we showed that $k^*(n) \geq \alpha_n bn/3$. This implies that $\frac{k^*(n)}{n} = \Omega(\alpha_n)$.

Part 3: Suppose that Σ_n is benign. Then because $R_k(\Sigma_n) \leq p_n - k$, we must have $p_n = \omega(n)$. Thus, we can restrict our attention to the sequences for which $p_n = \omega(n)$ and find the necessary and sufficient conditions for that class.

Next, for any positive α and any natural number $k \in [1, p_n)$, we can write

$$\int_k^{p_n} x^{-\alpha}\, dx \geq \sum_{i=k+1}^{p_n} i^{-\alpha} \geq \int_{k+1}^{p_n} x^{-\alpha}\, dx,$$

$$F(p_n) - F(k) \geq \sum_{i=k+1}^{p_n} i^{-\alpha} \geq F(p_n) - F(k+1),$$

where

$$F(x) = \begin{cases} \frac{1}{1-\alpha} x^{1-\alpha}, & \text{for } \alpha \neq 1, \\ \ln(x), & \text{for } \alpha = 1. \end{cases}$$

As the sequence can only be benign if $k^* = o(n)$, we can only consider values of k that do not exceed some constant fraction of n, e.g. $n/2$. Since $p_n = \omega(n)$, noting that, for $x > 0$, the sign of $\frac{1}{1-\alpha} x^{1-\alpha}$ flips when α crosses 1, we can write, uniformly for all $k \in [1, n/2]$,

$$\sum_{i=k+1}^{p_n} i^{-\alpha} = \begin{cases} \Theta_\alpha\left(p_n^{1-\alpha}\right), & \text{for } \alpha \in (0,1), \\ \Theta_\alpha\left(\ln(p_n/k)\right), & \text{for } \alpha = 1, \\ \Theta_\alpha\left(k^{1-\alpha}\right), & \text{for } \alpha > 1. \end{cases}$$

Recall that we consider $\lambda_{i,n} = i^{-\alpha}$ for $i \leq p_n$. Using the formula above, we get uniformly for all $k \in [1, n/2]$

$$r_k(\Sigma_n) = \begin{cases} \Theta_\alpha\left(k^\alpha p_n^{1-\alpha}\right), & \text{for } \alpha \in (0,1), \\ \Theta_\alpha\left(k \ln(p_n/k)\right), & \text{for } \alpha = 1, \\ \Theta_\alpha\left(k\right), & \text{for } \alpha > 1. \end{cases}$$

Recall that $k^* = \min\{k : r_k(\Sigma_n) \geq bn\}$. We compute

$$k^* = \begin{cases} \Theta_\alpha\left(p_n^{1-\frac{1}{\alpha}} n^{\frac{1}{\alpha}}\right), & \text{for } \alpha \in (0,1), \\ \Theta_\alpha\left(\frac{n}{\ln(p_n/n)}\right), & \text{for } \alpha = 1, \\ \Theta_\alpha\left(n\right), & \text{for } \alpha > 1. \end{cases}$$

One can see that for $\alpha > 1$, $k^* = \Omega_\alpha(n)$, so the sequence is not benign for $\alpha > 1$. On the other hand, $k^* = o(n)$ for $\alpha \leq 1$.

Next, analogously to the asymptotics for $r_k(\Sigma)$, we have

$$r_k(\Sigma_n^2) = \begin{cases} \Theta_\alpha\left(k^{2\alpha} p_n^{1-2\alpha}\right), & \text{for } \alpha \in (0, 0.5), \\ \Theta_\alpha\left(k \ln(p_n/k)\right), & \text{for } \alpha = 0.5, \\ \Theta_\alpha\left(k\right), & \text{for } \alpha \in (0.5, 1]. \end{cases}$$

Since $R_k = \frac{r_k(\Sigma)^2}{r_k(\Sigma^2)}$, we can write uniformly for all $k \in [1, n/2]$

$$R_k = \begin{cases} \Theta_\alpha\left(p_n\right), & \text{for } \alpha \in (0, 0.5), \\ \Theta_\alpha\left(\frac{p_n}{\ln(p_n/k)}\right), & \text{for } \alpha = 0.5, \\ \Theta_\alpha\left(k^{2\alpha-1} p_n^{2-2\alpha}\right), & \text{for } \alpha \in (0.5, 1), \\ \Theta_\alpha\left(\ln(p_n/k)^2\right), & \text{for } \alpha = 1. \end{cases}$$

Now we plug in k^* instead of k. Recall that $p_n/k^* = \Theta_\alpha\left((p_n/n)^{1/\alpha}\right)$ for $\alpha \in (0,1)$, and $p_n/k^* = \Theta_\alpha\left(p_n/n \ln(p_n/n)\right)$ for $\alpha = 1$. We get

$$R_{k^*} = \begin{cases} \Theta_\alpha\left(p_n\right), & \text{for } \alpha \in (0, 0.5), \\ \Theta_\alpha\left(n \frac{p_n/n}{\ln(p_n/n)}\right), & \text{for } \alpha = 0.5, \\ \Theta_\alpha\left(n \left(\frac{p_n}{n}\right)^{\frac{1}{\alpha}-1}\right), & \text{for } \alpha \in (0.5, 1), \\ \Theta_\alpha\left(\ln(p_n/n)^2\right), & \text{for } \alpha = 1. \end{cases}$$

Since $p_n = \omega(n)$, for any $\alpha \in (0,1)$, $R_{k^*} = \omega(n)$. For $\alpha = 1$ the necessary and sufficient for $R_{k^*} = \omega(n)$ is $\ln(p_n/n) = \omega(\sqrt{n})$.

So far, we obtained the necessary and sufficient conditions for the last terms to go to zero. Now let's look at the upper bound for the first term. We write, for $\alpha \in (0, 1]$,

$$r_0 = \sum_{i=1}^{p_n} i^{-\alpha} = \begin{cases} \Theta_\alpha\left(p_n^{1-\alpha}\right), & \text{for } \alpha \in (0,1), \\ \Theta_\alpha\left(\ln p_n\right), & \text{for } \alpha = 1. \end{cases}$$

Thus, for $\alpha < 1$, $r_0(\Sigma_n)/n$ goes to zero if and only if $p_n = o\left(n^{1/(1-\alpha)}\right)$, and for $\alpha = 1$, $r_0(\Sigma_n)/n$ goes to zero if and only if $\ln(p_n) = o(n)$.

Part 4: Suppose that Σ_n is benign. Then because $R_k(\Sigma_n) \le p_n - k$, we must have $p_n = \omega(n)$. Also,

$$\begin{aligned} \operatorname{tr}(\Sigma_n) &= \Theta\left(1 - e^{-p_n/\tau} + p_n \epsilon_n\right) \\ &= \Theta\left(1 + p_n \epsilon_n\right), \end{aligned}$$

and so $p_n \epsilon_n = o(n)$. Since Σ_n benign implies $k^* = o(n)$, and hence $k^* = o(p_n)$, we consider $k = o(p_n)$. In this regime,

$$\begin{aligned} \sum_{i>k} \lambda_i &= \Theta\left(e^{-k/\tau} - e^{-p_n/\tau} + (p_n - k)\epsilon_n\right) \\ &\le \Theta\left(e^{-k/\tau} + p_n \epsilon_n\right). \end{aligned}$$

Thus, whenever $k \le p_n$,

$$r_k(\Sigma_n) \le \Theta\left(\frac{e^{-k/\tau} + p_n \epsilon_n}{e^{-k/\tau} + \epsilon_n}\right).$$

Notice that

$$\frac{d}{dx} \frac{x + p_n \epsilon_n}{x + \epsilon_n} = \frac{\epsilon_n - p_n \epsilon_n}{(x + \epsilon_n)^2} < 0,$$

so k^* must be large enough to make

$$\frac{e^{-k/\tau} + p_n \epsilon_n}{e^{-k/\tau} + \epsilon_n} = \Omega(n).$$

Substituting $k = \tau \ln(n/(p_n \epsilon_n)) - a$ gives

$$\begin{aligned} r_k(\Sigma_n) &\le \Theta\left(\frac{p_n \epsilon_n / n + p_n \epsilon_n}{p_n \epsilon_n / n + \epsilon_n}\right) \\ &= \Theta\left(\frac{p_n \epsilon_n}{p_n \epsilon_n / n}\right) \\ &= \Theta\left(n\right), \end{aligned}$$

which shows that $k^* \geq \tau \ln(n/(p_n\epsilon_n)) - O(1)$. Thus, if Σ_n is benign, we must have $k^* = o(n)$, that is, $\epsilon_n p_n = ne^{-o(n)}$.

Conversely, assume $p_n = \Omega(n)$ and $\epsilon_n p_n = ne^{-o(n)}$ (that is, $\ln(n/(p_n\epsilon_n)) = o(n)$). Set $k = \tau \ln(n/(p_n\epsilon_n)) - a$, for some a, which we shall see is $\Theta(1)$. Notice that $k = o(n)$, so $p_n - k = \Omega(p_n)$ and $e^{-p_n} = o(e^{-k})$. Thus,

$$\begin{aligned}
\sum_{i>k} \lambda_i &= \Theta\left(e^{-k/\tau} - e^{-p_n/\tau} + (p_n - k)\epsilon_n\right) \\
&= \Theta\left(e^{-k/\tau} + p_n\epsilon_n\right), \\
\sum_{i>k} \lambda_i^2 &= \Theta\left(e^{-2k/\tau} - e^{-2p_n} + (p_n - k)\epsilon_n^2\right) \\
&= \Theta\left(e^{-2k/\tau} + p_n\epsilon_n^2\right).
\end{aligned}$$

These imply

$$\begin{aligned}
\mathrm{tr}(\Sigma_n) &= \Theta(1 + p_n\epsilon_n), \\
r_k(\Sigma_n) &= \Theta\left(\frac{e^{-k/\tau} + p_n\epsilon_n}{e^{-k/\tau} + \epsilon_n}\right) \\
&= \Theta\left(\frac{ap_n\epsilon_n/n + p_n\epsilon_n}{ap_n\epsilon_n/n + \epsilon_n}\right) \\
&= \Theta\left(\frac{p_n\epsilon_n}{ap_n\epsilon_n/n}\right) \\
&= \Theta\left(n/a\right),
\end{aligned}$$

which shows that $k^* = \tau \ln(n/(p_n\epsilon_n)) + O(1)$. Also, we have

$$\begin{aligned}
R_k(\Sigma_n) &= \Theta\left(\frac{\left(e^{-k/\tau} + p_n\epsilon_n\right)^2}{e^{-2k/\tau} + p_n\epsilon_n^2}\right) \\
&= \Theta\left(\frac{\left(p_n\epsilon_n/n + p_n\epsilon_n\right)^2}{p_n^2\epsilon_n^2/n^2 + p_n\epsilon_n^2}\right) \\
&= \Theta\left(\frac{p_n^2\epsilon_n^2}{p_n^2\epsilon_n^2/n^2 + p_n\epsilon_n^2}\right) \\
&= \Theta\left(\min\left\{n^2, p_n\right\}\right).
\end{aligned}$$

Combining gives

$$R(\hat{\theta}) = O\left(\frac{\epsilon_n p_n + 1}{n} + \frac{\ln(n/(\epsilon_n p_n))}{n} + \max\left\{\frac{1}{n}, \frac{n}{p_n}\right\}\right).$$

Now, it is clear that $p_n = \omega(n)$, $\epsilon_n p_n = o(n)$, and $\epsilon_n p_n = ne^{-o(n)}$ imply that Σ_n is benign.

Part 5: For $k < s_n$,

$$r_k(\Sigma_n) = (s_n - k) + \epsilon_n(p_n - s_n).$$

All the eigenvalues beyond the s_nth are the same. Thus, for $k \geq s_n$, we have $r_k(\Sigma_n) = R_k(\Sigma_n) = p_n - k$.

Now, for $n/R_{k^*}(\Sigma_n) \to 0$, we need $p_n = \omega(n)$, so $r_0(\Sigma_n)/n$ goes to 0 iff $s_n = o(n)$ and $\epsilon_n p_n = o(n)$. Also, $k^* = s_n$, so k^*/n goes to zero iff $s_n = o(n)$. Notice also that $p_n = \omega(n)$ and $s_n = o(n)$ imply that $n/R_{k^*}(\Sigma_n)$ also goes to zero. Thus, Σ_n is benign iff

$$s_n = o(n), \qquad \epsilon_n p_n = o(n), \qquad p_n = \omega(n).$$

Part 6: Notice that, for $1 < k \leq p_n$,

$$(1-\theta)^2 < 1 + \theta^2 - 2\theta\cos(k\pi/(p_n+1)) < (1+\theta)^2,$$

and hence

$$r_k \geq \frac{(p_n - k)(1-\theta)^2}{(1+\theta)^2}, \qquad\qquad R_k \geq \frac{(p_n - k)(1+\theta)^4}{(1-\theta)^4}.$$

Also,

$$\frac{(p_n - 1)(1-\theta)^2\epsilon_n}{(1+\theta)^2} + 1 \leq r_0 \leq \frac{(p_n - 1)(1+\theta)^2\epsilon_n}{(1-\theta)^2} + 1.$$

Thus, $r_0/n \to 0$ iff $p_n\epsilon_n = o(n)$. For $n/R_{k^*}(\Sigma_n) \to 0$, we need $p_n = \omega(n)$. Conversely, if $p_n = \omega(n)$, $k^* = 1$, and then $k^*/n \to 0$ and $n/R_{k^*}(\Sigma_n) \to 0$.

J. Upper bound on the B term. We can control the term ${\theta^*}^\top B\theta^*$ in Lemma 2 using a standard argument.

Lemma S.18. *There is a constant c, that depends only on σ_x, such that for any $1 < t < n$, with probability at least $1 - e^{-t}$,*

$${\theta^*}^\top B\theta^* \le c\|\theta^*\|^2\|\Sigma\| \max\left\{\sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{t}{n}}\right\}.$$

Proof. Note that

$$\left(I - X^\top \left(XX^\top\right)^{-1} X\right) X^\top = X^\top - X^\top \left(XX^\top\right)^{-1} (XX^\top) = 0. \tag{S3}$$

Moreover, for any v in the orthogonal complement to the span of the columns of $X^\top$,

$$\left(I - X^\top \left(XX^\top\right)^{-1} X\right) v = v.$$

Thus,

$$\|I - X^\top \left(XX^\top\right)^{-1} X\| \le 1. \tag{S4}$$

Now we can apply Eq. [S3] to write

$$\begin{aligned}
{\theta^*}^\top B\theta^* &= {\theta^*}^\top \left(I - X^\top \left(XX^\top\right)^{-1} X\right) \Sigma \left(I - X^\top \left(XX^\top\right)^{-1} X\right) \theta^* \\
&= {\theta^*}^\top \left(I - X^\top \left(XX^\top\right)^{-1} X\right) \left(\Sigma - \frac{1}{n} X^\top X\right) \left(I - X^\top \left(XX^\top\right)^{-1} X\right) \theta^*.
\end{aligned}$$

Combining with Eq. [S4] shows that

$${\theta^*}^\top B\theta^* \le \left\|\Sigma - \frac{1}{n} X^\top X\right\| \|\theta^*\|^2.$$

Thus, due to Theorem 9 in (4), there is an absolute constant c such that for any $t > 1$ with probability at least $1 - e^{-t}$,

$${\theta^*}^\top B\theta^* \le c\|\theta^*\|^2\|\Sigma\| \max\left\{\sqrt{\frac{r(\Sigma)}{n}}, \frac{r(\Sigma)}{n}, \sqrt{\frac{t}{n}}, \frac{t}{n}\right\},$$

where

$$r(\Sigma) := \frac{(\mathbb{E}\|x\|)^2}{\|\Sigma\|} \le \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|} = \frac{1}{\lambda_1} \sum_i \lambda_i = r_0(\Sigma).$$

□

K. Another lower bound. In this section, we prove the second paragraph of Theorem 1.

First, note that, without loss of generality, $||\Sigma||_2 = 1$ and $||\theta^*|| = 1$, since scaling these scales the excess risk by $||\Sigma||_2$ and $||\theta^*||^2$ respectively. This implies $\lambda_1 = 1$, and we may further assume without loss of generality that $\Sigma = \mathrm{diag}(\lambda_1, \lambda_2, \ldots)$. Define $s = \sum_{i=1}^\infty \lambda_i$. We may also assume that

$$\frac{r_0(\Sigma)}{n \log\left(1 + r_0(\Sigma)\right)} \ge c_2 \tag{S5}$$

since, otherwise, the lower bound is vacuously satisfied.

Define a metric ρ over $\mathbb{H}$ by

$$\rho(u, v) = \sqrt{(u - v)^\top \Sigma (u - v)},$$

so that, informally, a successful learning algorithm achieves $\rho(\hat{\theta}, \theta) < \sqrt{\tau_0}$.

Definition S.19. *Define sets $S_1, S_2, \ldots$ of indices as follows. Let $S_1 = \{1\}$; let $S_2 = \{2, \ldots, i_2\}$, for the least i_2 such that $\sum_{i=2}^{i_2} \lambda_i \ge 1$. Continue the same way as long as possible; for all $j > 2$, let $S_j = \{i_{j-1}, \ldots, i_j\}$, where i_j is the least index such that $\sum_{i=i_{j-1}}^{i_j} \lambda_i \ge 1$.*

Lemma S.20. *Definition S.19 produces $\Omega(n \log n)$ sets.*

Proof. For all j, $\sum_{i \in S_j} \lambda_i < 2$. Thus, for all k, $\sum_{i \le i_k} \lambda_i = \sum_{j \le k} \sum_{i \in S_j} \lambda_i < 2k$. Assume for contradiction that, for $k < \frac{c_2 n \ln n}{4}$, after S_k, it is not possible to add any more sets. Then $\sum_{i \le i_k} \lambda_i < \frac{c_2 n \ln n}{2}$, and, since no more sets can be added, $\sum_{i=1}^\infty \lambda_i < 1 + \frac{c_2 n \ln n}{2}$. We claim that, for large enough n, this contradicts the assumption that $\frac{\sum_{i=1}^\infty \lambda_i}{\ln\left(1 + \sum_{i=1}^\infty \lambda_i\right)} \ge c_2 n$. To see

why, consider the function $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ defined by $\phi(s) = \frac{s}{\ln(1+s)}$. The function ϕ is increasing for $s \geq 1$, so it suffices to show that $\phi\left(1 + \frac{c_2 n \ln n}{2}\right) < c_2 n$, and

$$\begin{aligned}\phi\left(1 + \frac{c_2 n \ln n}{2}\right) &= \frac{1 + \frac{c_2 n \ln n}{2}}{\ln\left(2 + \frac{c_2 n \ln n}{2}\right)} \\ &= \left(\frac{c_2}{2} + o(1)\right) n,\end{aligned}$$

yielding the contradiction and completing the proof. □

Definition S.21. *If the number of sets produced by the process of Definition S.19 is finite, let d be this finite number. Otherwise, let $d = \lceil n \ln n \rceil$.*

Now, informally, we, in our role as an adversary, commit to assigning all covariates in S_j the same weight. The following definition formalizes this idea.

Definition S.22. *Define a mapping ϕ from $\mathbb{R}^d$ to $\mathbb{H}$ as follows. For $w \in \mathbb{R}^d$, $\phi(w) = \theta$ where, for all j in $[d]$, for all i in S_j, $\theta_i = w_j$. For all $i > i_d$, $\theta_i = 0$.*

We would like to show that applying ϕ to an L_2 packing yields a ρ-packing, which is done in the following lemma.

Lemma S.23. *For all $u, v \in \mathbb{R}^n$, $\rho(\phi(u), \phi(v)) \geq ||u - v||$.*

Proof.

$$\begin{aligned}\rho(\phi(u), \phi(v))^2 &= \sum_i \lambda_i (\phi(u)_i - \phi(v)_i)^2 \\ &= \sum_j \left(\sum_{i \in S_j} \lambda_i\right) (u_j - v_j)^2 \\ &\geq \sum_j (u_j - v_j)^2.\end{aligned}$$

□

Let A be the least-norm interpolation algorithm. We will bound the accuracy of A by bounding its performance in terms of an algorithm C built using A as a subroutine, as was done in a related context in (10). The definition of Algorithm C is illustrated in Figure S1, which is reproduced from (10). The definition uses the function Q_α that rounds its input to the

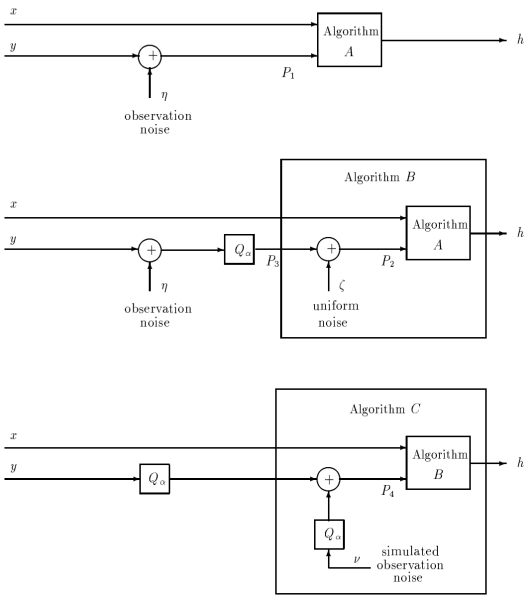

Fig. S1. A diagram illustrating the definition of Algorithm C.

nearest multiple of α. Algorithm C applies algorithm A to training data whose response variables have been modified. For

each example (x, y), and simulated artificial noise ε distributed as $N(0,1)$, and artificial noise ζ distributed uniformly on $(-\alpha/2, \alpha/2)$, Algorithm C gives $(x, y + Q_\alpha(\varepsilon) + \zeta)$ to A. The following lemma is similar to Lemma 5 of (10). One important difference is that we show that Algorithm C approximates the linear function parameterized by θ^*, not its discretization.

Lemma S.24. *If the linear interpolant algorithm A has error τ from n examples drawn from $N(0,\Sigma)$ with independent $N(0,1)$ noise with probability $1-\delta$, and*

$$\alpha \leq \min\left\{\frac{\delta}{2n}, 2\tau\right\}$$

then, in the absence of noise, Algorithm C, given n examples of the form $(x, Q_\alpha(\theta^\top x))$, with probability $1-2\delta$, achieves $\rho(\hat{\theta}, \theta^*)^2 \leq \tau$.

The proof of Lemma S.24 will be deferred until we have proved some more lemmas.

Recall the definition of total variation distance, $d_{TV}(P,Q) = \sup_E |P(E) - Q(E)|$. The following lemma is implicit in the proof of Lemma 6 of (10).

Lemma S.25. *Let η, ν be random variables that are distributed according to $N(0,1)$ and let ζ be uniform over $[-\alpha/2, \alpha/2]$.*

(a) For any $y \in \mathbb{R}$, if P_1 is the distribution of $y+\eta$ and P_2 is the distribution of $Q_\alpha(y+\eta)+\zeta$, we have $d_{TV}(P_1, P_2) \leq \alpha$.

(b) For any $y \in \mathbb{R}$, if P_3 is the distribution of $Q_\alpha(y+\eta)$ and P_4 is the distribution of $Q_\alpha(y) + Q_\alpha(\eta)$, $d_{TV}(P_3, P_4) \leq \alpha$.

We will use the following, which is implicit in the proof of Lemma 8 of (11).

Lemma S.26. *If $P_1, ..., P_n, Q_1, ..., Q_n$ are probability distributions over a domain U, and χ is a $[0,1]$-valued random variable defined on U^n then*

$$\left|\mathbb{E}_{\prod_t P_t}(\chi) - \mathbb{E}_{\prod_t Q_t}(\chi)\right| \leq \sum_{t=1}^n d_{TV}(P_t, Q_t).$$

Now, we are ready to prove Lemma S.24. The proof closely follows the proof of Lemma 5 in (10).

Proof (of Lemma S.24). Let $A(X, \boldsymbol{\varepsilon}, \theta^*)$ be the output $\hat{\theta}$ of the least-norm interpolant when the covariates are X, the noise is $\boldsymbol{\varepsilon}$, and the target is θ^*. Let $A(X, \boldsymbol{y})$ be the output $\hat{\theta}$ of the least-norm interpolant when the covariates are X, and the response variables are $\boldsymbol{y}$, and let $\mathrm{sam}(X, \boldsymbol{\varepsilon}, \theta^*) = (X, X\theta^* + \boldsymbol{\varepsilon})$ be the input arising from covariates X, regressor θ^* and noise $\boldsymbol{\varepsilon}$.

By assumption

$$N(0,\Sigma)^n \times N(0,1)^n\{(X, \boldsymbol{\varepsilon}) : \rho(A(\mathrm{sam}(X, \boldsymbol{\varepsilon}, \theta^*)), \theta^*)^2 \geq \tau\} < \delta.$$

Let ζ_t be a random variable with distribution U_α, where U_α is the uniform distribution over $(-\alpha/2, \alpha/2)$. Let B be the randomized algorithm that adds noise ζ_t to each y_t value it receives, passes the result to Algorithm A, and returns A's output.

Fix X, and define

$$E = \{\boldsymbol{\varepsilon} \in \mathbb{R}^n : \rho(A(\mathrm{sam}(X, \boldsymbol{\varepsilon}, \theta^*)), \theta^*)^2 \geq \tau\}$$
$$E_1 = \{\boldsymbol{y} \in \mathbb{R}^n : \rho(A(X, \boldsymbol{y}), \theta^*)^2 \geq \tau\}.$$

We have

$$N(0,1)^n(E) = \left(\prod_{t=1}^n P_{1|x_t}\right)(E_1),$$

where $P_{1|x_t}$ is the distribution of $(\theta^*)^\top x_t + \varepsilon_t$.

Define $P_{2|x_t}$ as the distribution of $Q_\alpha((\theta^*)^\top x_t + \varepsilon_t) + \zeta_t$. From Lemma S.25, $d_{TV}(P_{1|x_t}, P_{2|x_t}) \leq \alpha$. Applying Lemma S.26 with χ as the indicator function for E_1,

$$\left|\left(\prod_t P_{2|x_t}\right)(E_1) - \left(\prod_t P_{1|x_t}\right)(E_1)\right| \leq \alpha n.$$

Since $\alpha \leq \frac{\delta}{2n}$, this implies

$$\left(\prod_t P_{2|x_t}\right)(E_1) \leq \left(\prod_t P_{1|x_t}\right)(E_1) + \delta/2 = N(0,1)^n(E) + \delta/2.$$

Let $P_{3|x_t}$ be the distribution of $Q_\alpha((\theta^*)^\top x_t + \varepsilon_t)$, and let

$$E_3 = \{(\boldsymbol{y}, \boldsymbol{\zeta}) \in \mathbb{R}^n \times \mathbb{R}^n : \rho(A(X, \boldsymbol{y} + \boldsymbol{\zeta}), \theta^*)^2 > \tau\}$$

so that

$$\left(\prod_t P_{2|x_t}\right)(E_1) = \left(\prod_t (P_{3|x_t} \times U_\alpha^n)\right)(E_3).$$

Let $P_{4|x_t}$ be the distribution of $Q_\alpha((\theta^*)^\top x_t) + Q_\alpha(\varepsilon_t)$. Applying Lemma S.26, we get

$$\left| \left(\prod_t (P_{3|x_t} \times U_\alpha^n) \right)(E_3) - \left(\prod_t (P_{4|x_t} \times U_\alpha^n) \right)(E_3) \right| \leq \sum_{t=1}^m d_{TV}(P_{3|x_t}, P_{4|x_t}).$$

From Lemma S.25, $d_{TV}(P_{3|x_t}, P_{4|x_t}) \leq \alpha$, so

$$\begin{aligned}
\left(\prod_t (P_{4|x_t} \times U_\alpha^n) \right)(E_3) &\leq \left(\prod_t (P_{3|x_t} \times U_\alpha^n) \right)(E_3) + \delta/2 \\
&= \left(\prod_t P_{2|x_t} \right)(E_1) + \delta/2 \\
&\leq N(0,1)^n(E) + \delta.
\end{aligned}$$

Averaging over the random choice of X, the probability, for $(X, \boldsymbol{\zeta}, \boldsymbol{\varepsilon})$ distributed as $N(0,\Sigma)^n \times U_\alpha^n \times N(0,1)^n$, that $\rho(A(X, Q_\alpha(X\theta^*) + Q_\alpha(\boldsymbol{\varepsilon}) + \boldsymbol{\zeta})), \theta^*)^2 > \tau$, is at most

$$(N(0,\Sigma)^n \times N(0,1)^n)\{(X,\boldsymbol{\varepsilon}) : \rho(A(\mathrm{sam}(X,\boldsymbol{\varepsilon},\theta^*),\theta^*)^2 > \tau\} + \delta \;\leq\; 2\delta.$$

But $A(X, Q_\alpha(X\theta^*) + Q_\alpha(\boldsymbol{\varepsilon}) + \boldsymbol{\zeta})$ is the output of the randomized algorithm C, so this completes the proof. □

So, informally, we have shown that if the least norm interpolant can learn unit length weight vectors with noise and $N(0,\Sigma)$ data, then there is an algorithm C than can learn from quantized data without noise. The next step is to lower bound the error of C.

Recall that we have fixed an n, that $s \stackrel{\text{def}}{=} \sum_{i=1}^\infty \lambda_i \geq cn$, and that $\Sigma = \mathrm{diag}(\lambda_1, \lambda_2, ...)$.

We will use the following, which is an immediate consequence of Corollary S.6.

Lemma S.27. *For each row x_t of X, and each $q > 1$,*

$$\Pr(||x_t|| > q\sqrt{s}) \leq \exp(-q^2/c).$$

The proof of the following lemma borrows heavily from (12).

Lemma S.28. *If $1/\alpha = O(n)$, there is a constant τ such that, for any regression algorithm C, for all large enough n, if C is given n examples of the form $(X, Q_\alpha(X\theta^*))$, if the rows of X are n independent draws from $N(0,\Sigma)$, with probability at least $1/2$, its output $\hat{\theta}$ satisfies $\rho(\hat{\theta}, \theta^*)^2 > \tau$.*

Proof. For $\tau > 0$ to be chosen later, assume for contradiction that, with probability $1/2$, $\rho(\hat{\theta}, \theta^*)^2 \leq \tau$. For an absolute constant c_3, let G be a set of $(1/\tau)^{c_3 d}$ members of the unit ball in $\mathbb{H}$ that are pairwise separated by $3\sqrt{\tau}$ w.r.t. ρ so that, for distinct members g, h of G, $\rho(g,h)^2 > 9\tau$.

For each $X \in \mathbb{R}^{n \times \infty}$, and each $\theta \in \mathbb{H}$, define

$$\phi(X,\theta) = \begin{cases} 1 & \text{if } \rho(C(X, Q_\alpha(X\theta)), \theta)^2 \leq \tau \\ 0 & \text{otherwise} \end{cases}$$

and define

$$S = \mathbb{E}_X \left[\sum_{\theta \in G} \phi(X,\theta) \right].$$

Our assumption about the learning ability of C implies that

$$S \geq |G|/2 = (1/\tau)^{c_3 d}/2. \qquad \text{[S6]}$$

For any $g, h \in G$ for which $Q_\alpha(Xg) = Q_\alpha(Xh)$, since $\rho(g,h) > 3\sqrt{\tau}$, it cannot be the case that both $\phi(X,g)$ and $\phi(X,h)$ are both 1. Thus, recalling that $x_1, ..., x_n$ are the rows of X, and that all elements of G have length at most 1, we have

$$\begin{aligned}
S &\leq \mathbb{E}_X(|\{Q_\alpha(Xg) : g \in G\}|) \\
&= \mathbb{E}_X(|\{Q_\alpha(Xg) : g \in G\}| \mathbb{1}_{\max_t ||x_t|| < \sqrt{s}}) + \sum_{i=1}^\infty \mathbb{E}_X(|\{Q_\alpha(Xg) : g \in G\}| \mathbb{1}_{\lfloor \max_t ||x_t||/s \rfloor = i}) \\
&\leq \left(\frac{c_4\sqrt{s}}{\alpha} \right)^n + \sum_{i=1}^\infty (i\sqrt{s}/\alpha)^n \times \Pr(\max_t ||x_t|| \geq i\sqrt{s}) \\
&\leq \left(\frac{c_4\sqrt{s}}{\alpha} \right)^n + \sum_{i=1}^\infty (i\sqrt{s}/\alpha)^n \times n e^{-i^2/c_5} \quad \text{(by Lemma S.27)} \\
&\leq c_6 n \left(\frac{c_4\sqrt{s}}{\alpha} \right)^n.
\end{aligned}$$

Since $1/\alpha = O(n)$

$$|\{Q_\alpha(Xg) : g \in G\}| \leq \exp(O(n \log(ns))) = \exp(O(n \log(nd)))$$

since $d = \Theta(s)$. Since $d = \Omega(n \log n)$, for large enough n and small enough τ, this contradicts Eq. [S6], completing the proof. □

Now we are ready to put everything together to prove the second paragraph of Theorem 1. By Lemma S.24, it suffices to prove that, for a small enough constant τ_0, if $1/\alpha = O(n)$, with probability 1/2, Algorithm C, given examples $(x, Q_\alpha(\theta^\top x))$, with probability 1/2, fails to achieves $\rho(\hat{\theta}, \theta^*)^2 \leq \tau_0$. By Lemma S.28, this is the case, completing the proof.

References

1. S Page, S Grünewälder, Ivanov-regularised least-squares estimators over large RKHSs and their interpolation spaces, (arXiv), Technical Report arXiv:1706.03678 [math.ST] (2017).
2. R Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge Series in Statistical and Probabilistic Mathematics. (Cambridge University Press), (2018).
3. R Vershynin, Introduction to the non-asymptotic analysis of random matrices in *Compressed Sensing: Theory and Applications*, eds. YC Eldar, G Kutyniok. (Cambridge University Press), p. 210–268 (2012).
4. V Koltchinskii, K Lounici, Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23**, 110–133 (2017).
5. M Rudelson, R Vershynin, Hanson-Wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.* **18**, 1–9 (2013).
6. M Rudelson, R Vershynin, Sampling from large matrices: an approach through geometric functional analysis. *J. ACM* **54**, 21 (2007).
7. GH Golub, CF Van Loan, *Matrix Computations*. (The Johns Hopkins University Press), (1996).
8. RH Shumway, DS Stoffer, *Time Series Analysis and Its Applications*. (Springer-Verlag, Berlin, Heidelberg), (2005).
9. S Noschese, L Pasquini, L Reichel, Tridiagonal toeplitz matrices: Properties and novel applications. *Numer. Linear Algebr. with Appl.* **20**, 302–326 (2013).
10. PL Bartlett, PM Long, RC Williamson, Fat-shattering and the learnability of real-valued functions. *J. Comput. Syst. Sci.* **52**, 434–452 (1996).
11. PL Bartlett, Learning with a slowly changing distribution in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. (ACM), pp. 243–252 (1992).
12. GM Benedek, A Itai, Learnability with respect to fixed distributions. *Theor. Comput. Sci.* **86**, 377–389 (1991).